The Stereoscopic Zoom

Sergi Pujades, Frédéric Devernay, Laurent Boiron, Rémi Ronfard

The Stereoscopic Zoom

Sergi Pujades[1,2], Frédéric Devernay[2], Laurent Boiron[3], and Rémi Ronfard[3]

[1]MPI for Intelligent Systems; Tübingen, Germany
[2]Univ. Grenoble Alpes, Inria, LIG; Grenoble, France
[3]Univ. Grenoble Alpes, Inria, LJK; Grenoble, France

Abstract

We study camera models to generate stereoscopic zoom shots, i.e. using very long focal length lenses. Stereoscopic images are usually generated with two cameras. However, we show that two cameras are unable to create compelling stereoscopic images for extreme focal length lenses. Inspired by the practitioners' use of the long focal length lenses we propose two different configurations: we "get closer" to the scene, or we create "perspective deformations". Both configurations are build upon state-of-the-art image-based rendering methods allowing the formal deduction of precise parameters of the cameras depending on the scene to be acquired. We present a proof of concept with the acquisition of a representative simplified scene. We discuss the advantages and drawbacks of each configuration.

1. Introduction

Now that technical progress has made 3D cinema and television a reality, artists should be able to explore new narratives, which take advantage of the optical illusion of depth from stereopsis in the storytelling.

"Zooming" means the change of a lens' focal length. Lenses are either "prime" (fixed focal length) or "zoom". Most lenses equipped with a long focal length are zooms, because the cameraman needs to adjust the focal length to create the desired image frame. Thus in the paper we (ab)use the word "zoom" to refer to a long focal length lens.

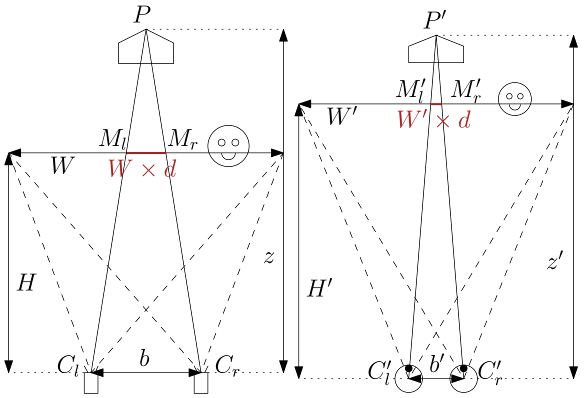
Zoom is one of the main limitations when shooting stereoscopic footage [25, 9]. The limitation arises from the incompatibility between the configuration avoiding ocular divergence, known to be a major cause of visual fatigue [40, 37], and the configuration avoiding the "cardboard effect", known to create a poor viewing experience [46, 25].

Yet zooms provide two opportunities to 2D cinematographers. The first is *to get closer* to the scene. Sometimes it is not possible to place the camera at the desired position, like for instance, when filming a polar bear in the wild. The second is aesthetic: it is well known that different focal lengths distort the perspective, and directors take advantage of these distortions to convey emotions. One of the most famous example in 2D is the *vertigo effect*, created by Alfred Hitchcock in 1958 in his feature film *Vertigo*. In stereoscopic movies too, directors should be given the opportunity to play with the perspective distortions at will to create new narratives yet to be invented.

Overview. In Sec.2 we introduce the notation and concepts related to stereoscopic filming. We illustrate the problems arising when filming with long focal lenses. In Sec. 3 we review the limitations of the existing state of the art methods. In Sec. 4 we present the simplified scene we use and in Sec. 5 and 6 we contribute two new camera models. One to *get closer* to the scene and one to *add perspective deformations* to the scene. In Sec. 7 we present a proof of concept for each camera model and in Sec. 8 we compare both configurations and conclude.

2. Problem Statement

Perceived Depth from Stereopsis. In Fig. 1 we introduce the parameters characterizing at the same time an *acquisition stereo system* and a *projection stereo system* [10]. In the acquisition setup, the distance between the optical centers of the camera b is the *baseline*. The cameras' *convergence distance* is H, and the plane parallel to the images at distance H is the *convergence plane*. The intersection of the camera visibility frustums with the convergence plane defines the convergence window. Its width is W. In the projection setup, the distance between the eyes of the spectator is b'. The distance between the spectator and the screen is H' and the width of the screen is W'. We assume b, b', H, H', W and W' are > 0. Given the acquisition parameters b, H, W, we can compute the normalized disparity d of an element at depth z, and given d and the projection



Symbol	Acquisition	Projection
$\mathbf{C}_l$, $\mathbf{C}_r$	camera optical center	eye optical center
$\mathbf{P}$	physical point of the scene	perceived 3D point
$\mathbf{M}_l$, $\mathbf{M}_r$	image points of $\mathbf{P}$	screen points
b	baseline	human eye distance
H	convergence distance	screen distance
W	convergence plane size	screen size
z	real depth	perceived depth
d	left-right disparity (as a fraction of W)	

Figure 1. Parameters describing the shooting and projection geometries (reproduced from [10]). (l and r indicate left and right).

parameters b', H', W', we can compute the perceived depth from stereopsis z':

$$d(z) = \frac{b}{W}\frac{(z-H)}{z} \text{ and } z'(d) = \frac{b'H'}{b' - W'd}. \qquad (1)$$

Combining both terms we can compute the relationship between the true depth in the 3D scene z and the perceived depth from stereopsis

$$z'(z) = \frac{zb'H'W}{z(b'W - bW') + bHW'}. \qquad (2)$$

Ocular Divergence Limits. Ocular divergence happens when both eyes look at the screen with a negative angle between them[1]. Both viewing rays intersect behind the spectator, i.e. $z' < 0$. The numerator of Eq. 2 can not be negative because z, b', W, H' are all positive. The denominator $bHW' + z(b'W - bW')$ is negative at $z = +\infty$ iif $b'W - bW' \leq 0$. The equality establishes the biggest non-divergence baseline:

$$b_{\text{div}} = b'\frac{W}{W'}. \qquad (3)$$

Note that if $b'W - W'b < 0$, only elements at depth $z > -\frac{bHW'}{(b'W - bW')}$ cause eye divergence in the projection room. Thus in a controlled acquisition setting, b_{div} could be slightly bigger.

Roundness Factor. In 1952 the *shape ratio* was defined as the ratio between depth magnification ($\frac{\partial z'}{\partial z}$) and width magnification of the perceived depth ($\frac{\partial x'}{\partial x}$ or $\frac{\partial y'}{\partial y}$) [40]. Later on, the term *roundness factor* has been used [25, 10]. Deriving Eq. 2 we obtain the expression of the roundness factor of an element at depth z

$$\rho(z) = \frac{bH'W}{z(b'W - bW') + bHW'}. \qquad (4)$$

If the roundness factor of a scene element is smaller than 0.3, it is perceived in depth, but it appears itself as flat, *as if it was drawn on a cutout cardboard* [25, 7]. Thus the roundness factor is important, as it allows to quantify the "cardboard effect". Establishing a target roundness for a specific depth (e.g. $z = H$), allows to compute the corresponding baseline:

$$b_{\rho(H)} = \frac{b'}{\rho(H)}\frac{H}{H'}. \qquad (5)$$

Long focal length lenses: ocular divergence vs. roundness. Let $f = \frac{H}{W}$ be the normalized acquisition focal length, and $f' = \frac{H'}{W'}$ the normalized projection focal length. The ratio between both focal lengths is

$$\frac{f}{f'} = \frac{b_{\rho(H)}}{b_{\text{div}}}\rho(H). \qquad (6)$$

Several studies study the best values for f', i.e. the optimal viewing distance with respect to the screen size [40, 2]. According to them it is reasonable to assume $f' \in [1.4, 2.5]$. However, actual long focal length lenses can easily reach normalized focal values $f = 10$[2]. Acquiring a stereoscopic pair of images with $f = 10$ and projecting them with $f' = 2.5$, will either create ocular divergence, or produce "cardboard effect".

Modifying the Perceived Depth. To adapt the content to different screen sizes, previous work propose to use a *disparity mapping function* $\phi(d) : \mathbb{R} \to \mathbb{R}$, transforming a disparity d into a mapped disparity $d' = \phi(d)$ [21, 11, 31]. We now study how a disparity mapping function affects the perceived depth from stereopsis. The disparity mapping function $\phi(d)$ is generally assumed to be increasing monotonic, to avoid mapping farther objects of the scene in front of nearer ones. Using $\phi(d)$ instead of d in Eq. 1 we obtain the *modified perceived depth from stereopsis*

$$z'(z) = \frac{b'H'}{b' - W'\,\phi\left(\frac{b}{W}\frac{(z-H)}{z}\right)}. \qquad (7)$$

In order to avoid ocular divergence

$$\phi(d) \leq \frac{b'}{W'} \quad \forall d \in \mathbb{R}, \qquad (8)$$

[1] A human can perform ocular divergence within a small range ($0.5 - 1°$) [37]. For the sake of simplicity we consider the limit to be $0°$.

[2] For instance the "Angenieux Optimo 28-340 cinema lens" [1]

which establishes the *mapped ocular divergence limits* constraint on $\phi(d)$. The *mapped roundness factor* is

$$\rho(z) = \frac{bH'}{z} \frac{\phi'(d(z))}{(b' - W'\phi(d(z)))}, \tag{9}$$

where ϕ' denotes the derivative of ϕ w.r.t. d. Note that $\forall d$, $\phi(d)$ needs to be differentiable.

3. Related Work

We review the domains of free-viewpoint video, blending multiple views, disparity mapping and multi-rigging.

Free-Viewpoint Video [39, 38] **.** The main idea of *free-viewpoint video* is that multiple images of a scene can be projected onto a geometric proxy in order to generate new realistic view-dependent images [18, 26, 23, 5, 41]. Most contributions in this domain target the productions of live events [13]. Stereoscopic images can be rendered from a standard camera configuration used for a 2D broadcast [15], i.e. combining their many cameras (up to 26). In that work authors do not explicitly address the long focal length shots, and our approach to the stereoscopic zoom could be integrated in such a framework.

Blending Multiple Views. Image-based rendering (IBR) is a plenoptic sampling problem [24]. One needs to reconstruct an optical ray from the available sampled rays. Unstructured Lumigraph Rendering [3] established the now prevailing [16, 8, 20] seven *desirable properties* that all IBR methods should fulfill. Among them, the *resolution sensitivity* property states: *"In reality, image pixels are not really measures of a single ray, but instead an integral over a set of rays subtending a small solid angle. This angular extent should ideally be accounted for by the rendering algorithm"* while the *minimal angular deviation* property states: *"source images rays with similar angles to the desired ray should be used when possible"*. Both can be formally deduced from the uncertainty of the 3D geometry [33]. The *minimal angular deviation* was also empirically devised to avoid visual artifacts [43]. In our work, we place the cameras using the *resolution sensitivity* and *minimal angular deviation* constraints.

Disparity Mapping Methods. As we saw in Sec. 2, a clever modification of disparity mapping function $\phi(d)$ can reduce the distortions in the 3D transformation between the acquired and perceived scenes, for instance avoiding ocular divergence or adding roundness. Disparity mapping functions can be linear [19], non linear [11] or a combination of disparity mapping operators [21, 31]. They are usually the same for all pixels in the image, but could be locally adapted to preserve details [42]. Once the disparity is modified, the novel view synthesis problem basically reduces to a view interpolation problem. We can classify the methods in two groups. First, *dense disparity maps warps* or Depth-Image-Based Rendering (DIBR) methods, generate a virtual viewpoint relying on a disparity map [34] using texture and depth information of the original images [48]. Second, *content aware warps* treat the novel view synthesis problem as a 2D mesh deformation problem, extensively studied in the field of media retargeting [44, 36, 14]. The idea is to consider the image as a regular grid, and compute the grid transformation preserving a set of constraints. In addition to stereoscopic, temporal and saliency constraints [21], constraints preserving lines and planes [47] can be used, as well as manually defined constraints [6, 22].

Content aware warps have two main limitations. First they only allows moderate modifications of the initial disparity – e.g. $\times 2 or \times 3$ expansion – in order to avoid visible stretch artifacts in the final images. Second it is unclear how to blend multiple images generated with these techniques, whereas blending is addressed in DIBR literature.

Multi-Rigging Techniques. In our approach we propose to use different cameras, each acquiring the scene with a different baseline, and then combine the images into the final shot (Sec. 5 and 6). Capturing one scene with several configurations is called *multi-rigging* [25, 10, 12]. The space is divided into depth regions, each acquired with a different configuration. Then the shots are composed depending on the depth of the elements. Special care is needed in the transition between configurations as visible artifacts could appear [30]. In live action stereoscopic 3D films, green screens are used to help with the depth composition [12], involving important human efforts. In CGI (computer-generated imagery) films, an "empty safe area" with no scene objects around the compositing depths allows avoiding visual artifacts [30].

Non-linear viewing rays or *bent rays* [30] can be used to smoothly transition between parts of the scene captured with different baselines. In fact, a multi-rig configuration can be associated with a disparity mapping function $\phi(d)$. In Sec. 6 we propose a multi-rig system and its associated $\phi(d)$. We propose to compute a world transformation Φ based on ϕ, so that we can handle the multi-rig problem as an IBR problem in a world where the optical rays are not straight.

4. Simplified Scene

To demonstrate the different possibilities to create a stereoscopic zoom, we focus on a simplified layout of the scene. It consists of a main subject and a background. It is a classic scenario where zooms are used in 2D, for instance to create a closeup of a soccer player focusing before a penalty kick. The physical cameras cannot disturb the performance, thus in our simplified scene we assume that the cameras cannot be "on the field". Figure 2 illustrates a simplified scene representing a player on the field with the bleachers on the background. Let z_s and z_b be respectively

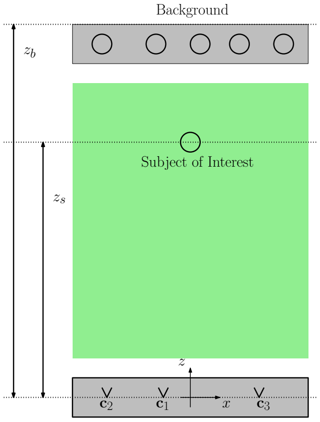


Figure 2. Simple scene with a subject of interest and a background. Actual cameras can only be placed outside the field. The distance between the actual camera location and the subject of interest and the background are z_s and z_b respectively.

the distance between the cameras and the subject of interest, and the distance between the cameras and the background. Our goal is to establish camera positions $\boldsymbol{c}_i$ and parameters to render the stereoscopic images following the director's mise-en-scene. We assume the virtual and actual cameras have the same image resolution.

5. Being On the Field

The Mise-en-Scene. The first stereoscopic mise-en-scene is the unconstrained placement of the cameras if it were possible: they would be on the field. The director freely defines a virtual filming configuration (b_v, H_v, W_v) in order to obtain the desired perceived depth z' with the projection parameters (b', H', W'). In Fig. 3 we illustrate a virtual configuration producing a linear depth mapping. The goal now is how to place the actual acquisition cameras in order to best render the virtual images.

The Quadri-Rig. To place the actual cameras, we use the *minimal angular deviation* and *resolution sensitivity* proposed by [3] and formalized by [33]. We place the actual cameras one by one and proceed in two stages: we first choose the focal length and then the camera position. Because the scene can be roughly decomposed in two layers, we propose to use two actual cameras to generate each virtual view. To generate the left virtual view we propose to use a camera to acquire the subject of interest and another one to acquire the background. Symmetrically, we use two cameras to generate the right virtual view. We name the resulting camera model the "Quadri-Rig".

Choosing the Focal Length. To obtain an image of a flat element with an equivalent resolution at a distance z_r with a focal length f_r and at a distance z_v with a focal length f_v, the relation between the focal lengths is $f_r = f_v \frac{z_r}{z_v}$. Note that this computation is only valid for a flat element at a single depth. Elements in front (or behind) this depth have an increasing (or decreasing) image size, depending on the distances z_r and z_v as shown in Fig. 4.

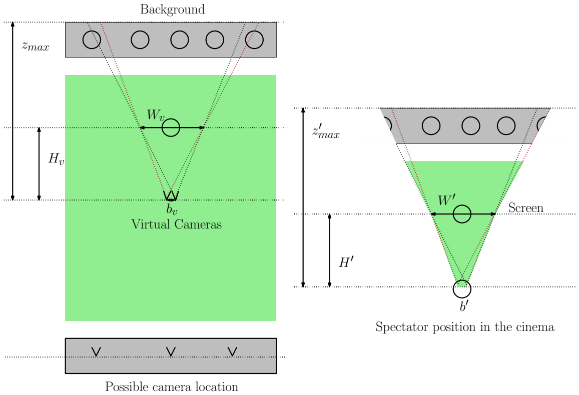


Figure 3. Acquisition and projection of the simplified scene. We illustrate a possible mise-en-scene. The director freely chooses the virtual cameras (left), so that the perceived depth from stereopsis in the projection room presents no distortions (right).

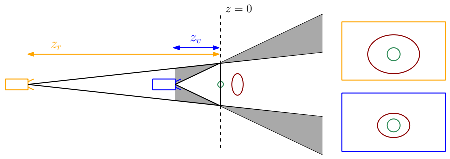


Figure 4. Two cameras with different focal lengths acquire a green object with the same resolution. The object has the same size on both images. A red object farther away has a bigger image size in the camera with a longer focal length.

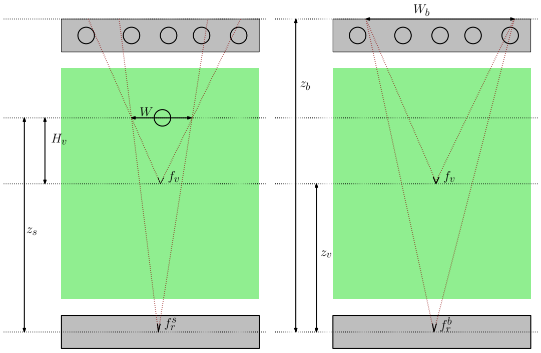


Figure 5. Acquiring a part of the scene with the same resolution as a virtual camera with focal length f_v. The focal length f_s acquires the subject of interest with the same resolution as the virtual camera. The focal length f_b acquires the background with the same resolution as the virtual camera.

By symmetry, both cameras acquiring the subject of interest have the same focal length f_s, and both cameras acquiring the background have the same focal length f_b (see Fig. 5):

$$f_s = f_v \frac{z_s}{H_v} \text{ and } f_b = f_v \frac{z_b}{z_b - z_v}. \tag{10}$$

Choosing the camera positions. Our camera position

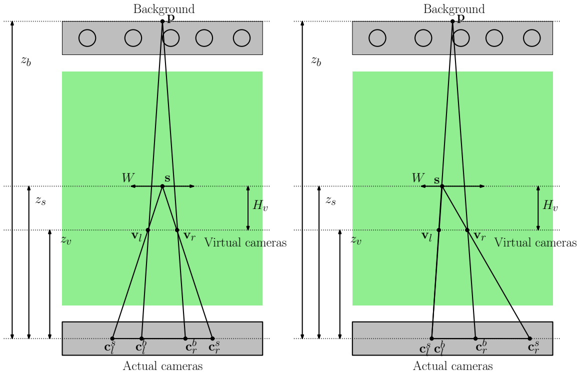


Figure 6. Diagram of the obtained camera positions for the "Quadri-Rig". Left: the cameras acquiring the subject $\boldsymbol{c}_*^s$ are aligned with the position of the subject of interest $\boldsymbol{s}$ and the virtual cameras $\boldsymbol{c}_*^v$. The cameras acquiring the background $\boldsymbol{c}_*^b$ are aligned with the center of the background $\boldsymbol{p}$ and the virtual cameras $\boldsymbol{c}_*^v$. Right: When $\boldsymbol{s}$, $\boldsymbol{p}$ and $\boldsymbol{v}_l$ are aligned, $\boldsymbol{c}_l^b$ and $\boldsymbol{c}_l^s$ are equal.

goal for the actual cameras can be stated as the position minimizing the *angular deviation* between the optical rays of the virtual camera and the actual camera capturing the scene element. To avoid the need of a geometric estimate of the acquired scene elements, we assume them to be punctual. Then the actual camera position minimizing the *angular deviation* is the one fulfilling the *epipolar consistency*: the camera is aligned with the optical center of the virtual camera and the position of the element to render (see Fig. 6). The remaining question is how to choose the point representing the acquired scene element. A natural choice seems to select the center of the subject as its simplified 3D position. The center of the subject of interest can be easily approximated as the center of gravity of the 3D subject's points seen by the camera. Similarly, the center of the background can be chosen as the center of the background seen by both images.

6. Distort the World!

We now propose a novel camera model inspired by the 2D-to-3D conversion methods, where the director establishes multiple depth and roundness constraints on an initial 2D frame (see Fig. 7). We study i) how they translate into multiple acquisition settings, and ii) how to combine the acquired images into the final stereoscopic effect.

The Mise-en-Scene. First the director chooses the 2D frame of the image, by placing a camera on the possible location area and adjusts the focal length to frame the subject of interest. Then, for each relevant element of the scene at depth z_e, the director specifies the expected perceived depth in the projection room $z'(z_e) = z'_e$. In addition, the director may also specify the expected *roundness factor* of each element, i.e. $\rho(z_e) = \rho_e$. Although at the time of the stereoscopic mise-en-scene the director is most probably unaware of the actual depth of the scene, the provided depth description establishes constraints on the perceived depth function $z'(z)$ (Eq. 2) and the roundness factor function $\rho(z)$ (Eq. 4).

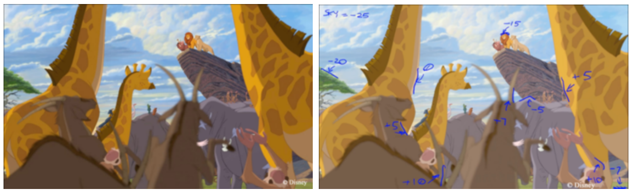


Figure 7. The original image (left) is annotated with the expected disparity in the final shot (right). Images reproduced from [28].

Our goal now is to translate the depth and roundness factor constraints into a (potentially) multi-view acquisition device. The 6 parameters of $z'(z)$ from Eq. 2 are (b, H, W) and (b', H', W'). Assuming the projection parameters fixed, then $z'(z)$ has only three degrees of freedom left. Specifying more than three constraints creates an over-determined system, i.e. one acquisition setting is not enough.

The first constraint on our setting is the focal length of the camera, chosen to create the 2D frame: $f = \frac{W}{H}$. Then, one depth constraint $z'(z_e) = z'_e$ together with a roundness constraint $\rho(z_e) = \rho_e$ fully constrain the acquisition setup. Similarly, if the convergence distance is kept constant for all constraints ($z'(H) = H'$) then only a depth or roundness constraint fixes the acquisition baseline. We may thus need as many acquisition cameras as constraints.

The Tri-Rig. For a generic scene containing many elements, it becomes unreasonable to use as many cameras as depth constraints. But a scene with a low number of constraints can be acquired with a small number of cameras. In our simplified scene layout with two elements, a constraint for the depth and roundness of the subject of interest, and a depth constraint on the background, result in a three camera configuration: *the Tri-Rig*.

We choose the camera set by the director to establish the 2D frame as the leftmost camera. Then, the second camera is chosen so that the subject of interest is perceived at the desired depth with the desired roundness ($z'(z_s) = z'_s, \rho(z_s) = \rho_s$). For example, choosing $z_s = H, z'(H) = H'$ and $\rho_s = 1$, defines the baseline between the left most camera and the second one $b_{\text{round}} = b' \frac{H}{H'}$. This configuration also establishes the convergence window width and distance (W, H) that we will keep fix. Then the third camera is placed so that the perceived depth of the background is the desired one ($z'(z_b) = z'_b$). For example, using $z'(\infty) = \infty$ to avoid ocular divergence gives $b_{\text{div}} = b' \frac{W}{W'}$.

Reference baseline and disparity mapping. We have now acquired three images and we need to compose them into a stereoscopic pair. The initial director's image is used as the leftmost image. Now we need to render a right image following the director's constraints. Notice that while in the "Being on the field" approach we had virtual cameras

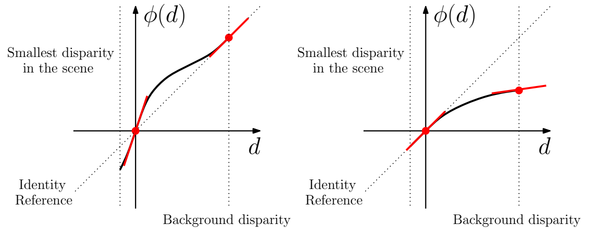


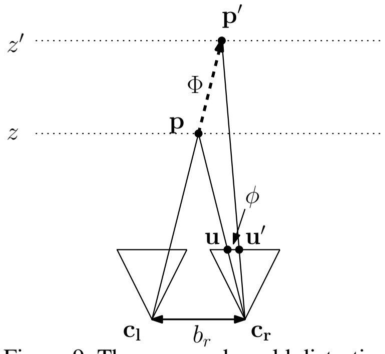
Figure 8. Disparity mapping function $\phi(d)$ examples. Left: $\phi(d)$ expanding the disparity range near the zero disparity, i.e. the subject of interest depth. Disparity values after the expansion are compressed and the disparity values of the background are preserved. Right: $\phi(d)$ compressing the disparity range for the background values. Disparity values at the convergence depth are preserved.

establishing the cameras to be rendered, in this case we do not know where the camera to be rendered is. Furthermore, the target image cannot be obtained with a standard pinhole camera. We thus describe the image formation process of the target view with the composition of two functions.

First, a function $d(z)$ transforms scene depth values z into disparity values d, and then a *disparity mapping* function $\phi(d)$ transforms the acquired disparities d into the desired ones. To define the depth to disparity function $d(z)$ we need a *reference* acquisition setup (b_r, H_r, W_r), which can be arbitrarily chosen among any director's constraint. Then, each supplementary constrain is taken into account by defining a *control point* on the $\phi(d)$ function. A depth constraint $z'(z_e) = z'_e$ constrains $\phi(d)$, whereas a roundness constraint $\rho(z_e) = \rho_e$ constrains $\phi(d)$ and $\phi'(d)$. Once all *control points* set, the final continuous function $\phi(d)$ can be computed with any interpolation technique exactly interpolating the control points and its derivatives [31, 21]. $\phi(d(z))$ must be differentiable, otherwise the roundness factor is not be defined. The shape of $\phi(d)$ varies depending on the *reference baseline*. In Fig. 8 we show two functions obtained using either b_{div} or b_{round} as reference.

Blending multiple images and the world distortion. Once ϕ is defined we can theoretically compute the warps from the input images into the target image and use [33] to render the images. Even though the blending weights of [33] can be computed as they only rely on the image transformation, it is unclear how to compute the weights for the method proposed by [3], as they rely on angles between optical rays which are affected by the $\boldsymbol{\phi}$ function.

Moreover, the actual implementation of the warps between the input and target images needs special attention, as the occlusion handling in this process is also affected by ϕ. Two geometric elements $\boldsymbol{p}_1$ and $\boldsymbol{p}_2$ projected at two different image locations $\boldsymbol{u}_1$ and $\boldsymbol{u}_2$, may have the same image coordinates after the disparity mapping warp. The visibility test, also known as z-buffering, should use the final disparity mapped values d' instead of the depth of the element to the camera z. The element with a lower disparity value occludes the element with a higher disparity value. Similarly, when computing the inverse warp, two image points $\boldsymbol{u}'_1$ and $\boldsymbol{u}'_2$, may be warped into points $\boldsymbol{u}_1$ and $\boldsymbol{u}_2$ having the same x- and y-coordinates. Thus, to first warp the image and then reconstruct the 3D points can not be done by storing the warped values in a classical single planar buffer. Some elements of the buffer will be empty, whereas other elements will have multiple assignments. A special pipeline should be implemented. However, as the occlusion handling in the render engines such as OpenGL [45] has been optimized over the years for standard pinhole camera projections, we would like to take advantage of the actual rendering techniques. Hence we propose not to apply the disparity mapping $\phi(d)$ in the images, but to distort the world accordingly with a function $\boldsymbol{\Phi} : \mathbb{R}^3 \rightarrow \mathbb{R}^3$ before the pinhole camera projection. With this pipeline we can compute the desired image warps with a classic depth occlusion handling, and, we will also be capable to compute angles between the viewing rays, as required by [3].



$$\begin{tikzcd}
\mathbf{p} \arrow[r, "\Phi(\mathbf{p})"] \arrow[d, "\tilde{\mathbf{P}}_r(\mathbf{p})"'] & \mathbf{p}' \arrow[d, "\tilde{\mathbf{P}}_r(\mathbf{p}')"] \\
\mathbf{u} \arrow[r, "\phi(\mathbf{u})"'] & \mathbf{u}'
\end{tikzcd}$$

Figure 9. The proposed world distortion Φ as the composition of: the projection of $\boldsymbol{p}$ to $\boldsymbol{u}$, the image warp on $\boldsymbol{u}$ defined by $\phi(d)$, and the backprojection of $\boldsymbol{u}$ into $\boldsymbol{p}'$. Left: Top view scheme of the scene. Right: function composition graph.

Originally the disparity mapping function was introduced to operate in image space, as the main applications targetted post-production [21, 11]. However, if we are at the acquisition stage, we can introduce a world deformation $\boldsymbol{\Phi}$ based on ϕ. In Fig. 9 we illustrate the construction of the proposed world distortion. Let us consider two cameras, the left at $\boldsymbol{c}_l$ and the right at $\boldsymbol{c}_r$, defining a reference setup (b_r, H_r, W_r). A point $\boldsymbol{p} = (x, y, z)$ in the original scene is projected into the right image point $\boldsymbol{u} = (u, v, d(z))$, where $d(z)$ is obtained by using (b_r, H_r, W_r) in Eq. 1. The image point $\boldsymbol{u}$ is then mapped into $\boldsymbol{u}' = (u', v, \phi(d(z)))$ using the predefined $\phi(d)$. Then a 3D point $\boldsymbol{p}' = (x', y', z')$ can be reconstructed as follows. It's z-coordinate is obtained as $z'(\phi(d(z)))$ using (b_r, H_r, W_r) in Eq. 1. Then we still need to define x' and y' to obtain the 3D distortion of the world $\boldsymbol{\Phi}$. As the frame of the left camera is chosen by the director, a natural constraint on $\boldsymbol{\Phi}(\boldsymbol{p})$ is that both $\boldsymbol{p}$ and $\boldsymbol{p}'$ project on the same point on the left camera. This way, the image

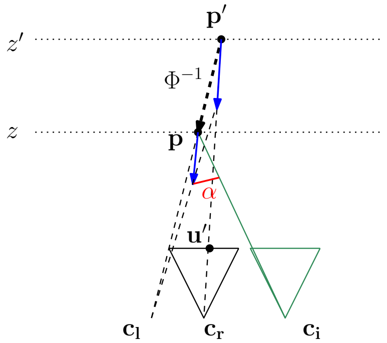


Figure 10. Computing angles between viewing rays to render the image point $\boldsymbol{u}'$ with the camera $\boldsymbol{c_i}$. The direction of the viewing ray defined by $\boldsymbol{u}'$ and $\boldsymbol{p}'$ is undistorted using $D\boldsymbol{\Phi}^{-1}$. The angle α between the undistorted ray and $\boldsymbol{p}$ and $\boldsymbol{c_i}$ can be computed.

of the left camera is unaffected by the geometry distortion, and $\boldsymbol{\Phi}(\boldsymbol{p}) = \boldsymbol{p}'$ is fully defined. Let us point out that if $\phi(d)$ is strictly increasing and differentiable, then $\boldsymbol{\Phi}(\boldsymbol{p})^{-1}$ exists and is well defined.

With the proposed world distortion, angles between the desired optical ray and the input camera ray can be computed as illustrated in Fig. 10. Given a point $\boldsymbol{u}'$ on the reference image, its 3D point $\boldsymbol{p}'$ in the distorted world can be computed. Then the desired viewing ray in the distorted world can be undistorted using the $D\boldsymbol{\Phi}^{-1}$ evaluated at the depth of $\boldsymbol{p}'$. The undistorted ray can now be compared to the ray between $\boldsymbol{p}$ and $\boldsymbol{c}_i$, where $\boldsymbol{p}$ is the undistorted version of $\boldsymbol{p}'$, and $\boldsymbol{c}_i$ the optical center of the input view. The weights of the method proposed by [3] can now be computed.

With the proposed world distortion we extend the disparity mapping problem into a more general image-based rendering problem. In practice $\boldsymbol{\Phi}(\boldsymbol{x})$ can be efficiently implemented with a vertex shader. While our approach distorts the world to obtain straight viewing rays to properly handle occlusions, the distorted world should not be used to compute any geometric values, such as angles or distances, as for instance, illumination techniques relying on them would lead erroneous results.

7. Proofs of Concept

We present proofs of concept of the proposed approaches on a synthetic dataset, *blender lego*, where the exact camera parameters and the exact geometry of the scene are known.

Quadri-Rig. To demonstrate the quadri-rig we rendered 6 images (see Fig.11). The left and right virtual images were rendered at the desired virtual positions $\boldsymbol{v}_l$ and $\boldsymbol{v}_r$. These images were used as ground truth for comparison with the rendered images. The other four images of the dataset correspond to the "Quadri-Rig" configuration: two images of the cameras acquiring the subject and two images of the cameras acquiring the background.

In Table 1 we present the PSNR and DSSIM computed values between the ground truth virtual images and the rendered ones using [3] and [33]. Because of the low number of images, all methods yield very similar results. This result is coherent with the results obtained in [32]. The difference in the blending weights has no significant impact on the rendered images when few images are used. The high PSNR and low DDSIM values obtained with the synthetic dataset show that the rendered images at visible locations are accurate. In the first row of Fig. 12 we reproduce the stereoscopic pair rendered with [33]. Because the background of the scene is purely flat, large regions visible in the virtual views are not acquired by any of the four actual cameras. The closeups in Fig. 12 show these occlusions.

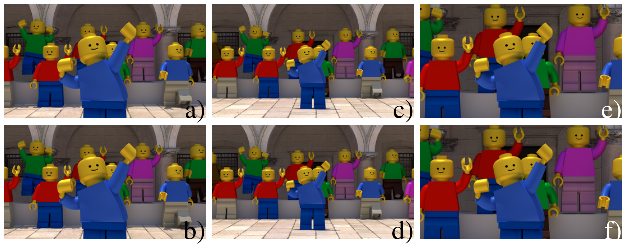


Figure 11. The "Quadri-Rig" *blender lego* dataset images. a) and b) is the virtual stereoscopic pair. c) and d) are the images acquired with the background cameras. e) and f) are the images acquired with the subject of interest cameras.

	left image		*right image*	
[3]	33.93	288	33.94	290
[33]	**34.00**	**287**	**34.02**	**288**

Table 1. Numerical results for the synthetic dataset. We compare [3] and [33]. The first value is PSNR (bigger is better), the second value is DSSIM in units of 10^{-4} (smaller is better). The best value is highlighted in bold.

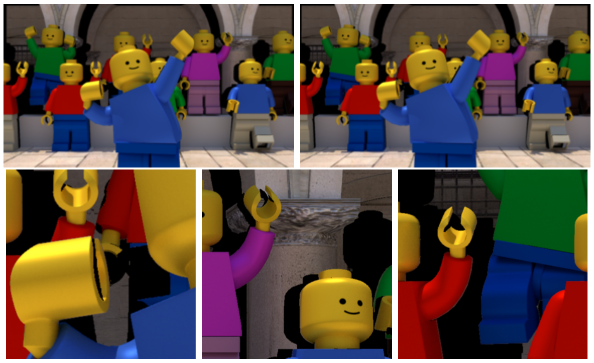


Figure 12. First row: stereoscopic pair obtained from the Quadri-Rig with [33]. Second row: closeups of the rendered views. As the background is not a plane, large areas in black are not acquired by any actual camera (compare with Fig. 15).

Tri-Rig. To demonstrate the "Tri-Rig" we rendered three views of the *blender lego* dataset (see Fig. 13) The used *reference baseline* is b_{round} and the disparity mapping function $\phi(d)$ has the shape illustrated in Fig. 8 right. Figure 14

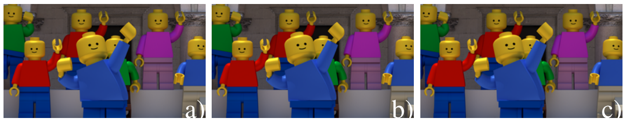


Figure 13. The “Tri-Rig” images of the *blender lego* dataset. a) and b): left and right images acquiring the background. a) and c): left and right images acquiring the subject of interest.

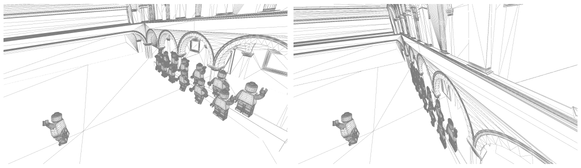


Figure 14. The world distortion for the “Tri-Rig” *blender lego* dataset. Left: original 3D scene. Right: Distorted 3D scene. The background depth is compressed, so that further elements create a smaller disparity in the final rendered image.

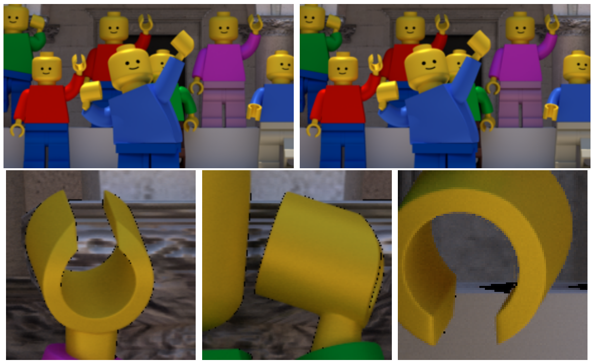


Figure 15. First row: Right images rendered from the Tri-Rig with [3] (left) and [33] (right). No noticeable difference is visible between the images. Second Row: Occluded regions in the “Tri-Rig”. Few pixels around depth discontinuities are not acquired by any camera (compare with Fig. 12).

illustrates the world distortion created by $\boldsymbol{\Phi}(\boldsymbol{x})$. The subject’s depth is preserved, while the background elements are pulled forward to decrease their disparity values on the final image. In the first row of Fig. 15 we show the final rendered images using [3] and [33], which, visually, do not present any noticeable difference. In the second row of Fig. 15 we show closeups of the regions in the rendered images, which are not acquired by any source image. These regions are very small compared to the large black areas obtained with the “Quadri-Rig” approach shown in Fig. 12. Moreover, let us recall that the left image of the target stereoscopic pair is the original one, i.e. it does not suffer at all of occluded regions. As the target image does not correspond to a perspective camera, we do not have a reference image to compare with, and thus we can not numerically evaluate the obtained results. Our approach distorts the world to obtain straight viewing rays and properly handle occlusions. However, the distorted world should not be used to compute any metric values, such as angles or distances. For instance, rendering techniques such as illumination, relying on metric values, should not be computed in the distorted world as they would lead to erroneous results.

8. Discussion and Conclusion

In this paper we addressed the problem of generating stereoscopic images with long focal lenses. We have illustrated why multiple cameras are needed and deduced two different camera models, the “Quadri-Rig” and “Tri-Rig”. Each model is inspired by the intention of the director, either to *get closer* to the scene or to add aesthetic *perspective deformation* to the scene. Although the mise-en-scene of each method is very different, we compare both camera models to highlight their advantages and flaws.

A key advantage of the “Tri-Rig” with respect to the “Quadri-Rig” is that one of the images, i.e. the left one, is acquired by a source camera. Indeed, the perceived quality of a stereoscopic pair of images is close (and sometimes equal) to the quality of the best of both images [35]. Thus having the raw output of the camera as the left view provides the highest quality possible. Moreover, as illustrated in Fig. 12, large areas needed by the target images of the “Quadri-Rig” may not be acquired by any of the four actual cameras. Although these regions could be filled in using an inpainting method [29, 17, 27, 4], these occlusion regions are smaller in the “Tri-Rig” setup, as shown in Fig. 15.

In this work the camera models target a scene with a simple layout. Nevertheless the proposed camera models are generic and can be extended to scenes with more elements. For each new relevant element in the scene, a new pair of cameras are needed. However, as in the “Tri-Rig” all cameras have the same focal length, all the left cameras of each configuration can be the same. Thus, given a scene with $N \geq 2$ relevant elements, the “Quadri-Rig”s complexity is $2N$, whereas the “Tri-Rig”s complexity is $N + 1$.

Because of the advantages of the “Tri-Rig” with respect to the “Quadri-Rig”, the director could also use the “Tri-Rig” with the intention to *get closer* to the scene. In this case, similarly to the use of a zoom in 2D, the perspective distortions would be a consequence, not the intention.

Future work should address the validation of the images generated by the “Tri-Rig”. As we do not have a reference image to compare with, we could not assess the relevance of the proposed camera model. We believe that a subjective evaluation of the obtained results should be conducted in the future to assess the proposed approach. One possibility to evaluate if the “Tri-Rig” is capable to create compelling stereoscopic images would be to conduct a user study. The observers would be presented with images generated with the “Tri-Rig” and images generated with disparity mapping methods using only two images [21, 47, 11]. The observer could then choose if one is preferred, or equal preference.

References

[1] Angenieux. Online angenieux portefolio. http://www.angenieux.com/zoom-lenses/cinema-portfolio/optimo-28-340.htm, 2015. [Online; accessed 14-August-2015]. 2

[2] M. S. Banks, J. Kim, and T. Shibata. Insight into vergence-accommodation mismatch. In *Head- and Helmet-Mounted Displays XVIII: Design and Applications*, pages 873509–873509–12. SPIE, 2013. 2

[3] C. Buehler, M. Bosse, L. McMillan, S. Gortler, and M. Cohen. Unstructured lumigraph rendering. In *SIGGRAPH*, pages 425–432. ACM, 2001. 3, 4, 6, 7, 8

[4] P. Buyssens, M. Daisy, D. Tschumperlé, and O. Lézoray. Depth-aware patch-based image disocclusion for virtual view synthesis. In *SIGGRAPH Asia 2015 Technical Briefs*, page 2. ACM, 2015. 8

[5] J. Carranza, C. Theobalt, M. A. Magnor, and H.-P. Seidel. Free-viewpoint video of human actors. In *SIGGRAPH*, pages 569–577. ACM, 2003. 3

[6] C.-H. Chang, C.-K. Liang, and Y.-Y. Chuang. Content-aware display adaptation and interactive editing for stereoscopic images. *Transactions on Multimedia*, 13(4):589–601, 2011. 3

[7] A. Chapiro, O. Diamanti, S. Poulakos, C. O'Sullivan, A. Smolic, and M. Gross. Perceptual evaluation of cardboarding in 3D content visualization. In *Proceedings of the ACM Symposium on Applied Perception*, SAP '14, pages 47–50, New York, NY, USA, 2014. ACM. 2

[8] G. Chaurasia, S. Duchene, O. Sorkine-Hornung, and G. Drettakis. Depth synthesis and local warps for plausible image-based navigation. *Transactions on Graphics*, 32(3):30, 2013. 3

[9] W. Chen. *Multidimensional characterization of quality of experience of stereoscopic 3D TV*. Theses, Université de Nantes Angers Le Mans, 2012. 1

[10] F. Devernay and P. Beardsley. Stereoscopic cinema. In R. Ronfard and G. Taubin, editors, *Image and Geometry Processing for 3D Cinematography*, pages 11–51. Springer Berlin Heidelberg, 2010. 1, 2, 3

[11] F. Devernay and S. Duchêne. New view synthesis for stereo cinema by hybrid disparity remapping. In *International Conference on Image Processing*, pages 5–8. IEEE, 2010. 2, 3, 6, 8

[12] C. Dsouza. *Think in 3D: Food For Thought for Directors, Cinematographers and Stereographers*. CreateSpace Independent Publishing Platform, 2012. 3

[13] M. Germann, T. Popa, R. Keiser, R. Ziegler, and M. Gross. Novel-view synthesis of outdoor sport events using an adaptive view-dependent geometry. *Computer Graphics Forum*, 31(2pt1):325–333, 2012. 3

[14] Y. Guo, F. Liu, J. Shi, Z.-H. Zhou, and M. Glei. Image retargeting using mesh parametrization. *Transactions on Multimedia*, 11(5):856–867, 2009. 3

[15] A. Hilton, J.-Y. Guillemaut, J. Kilner, O. Grau, and G. Thomas. 3D-TV production from conventional cameras for sports broadcast. *Transactions on Broadcasting*, 57(2):462–476, 2011. 3

[16] A. Hornung and L. Kobbelt. Interactive pixel-accurate free viewpoint rendering from images with silhouette aware sampling. *Computer Graphics Forum*, 28(8):2090–2103, 2009. 3

[17] V. Jantet, C. Guillemot, and L. Morin. Joint projection filling method for occlusion handling in depth-image-based rendering. *3D Research*, 2(4):1–13, 2011. 8

[18] T. Kanade, P. Rander, and P. Narayanan. Virtualized reality: constructing virtual worlds from real scenes. *MultiMedia*, 4(1):34–47, 1997. 3

[19] H. J. Kim, J. W. Choi, A.-J. Chang, and K. Y. Yu. Reconstruction of stereoscopic imagery for visual comfort. In *Stereoscopic Displays and Applications XIX*, pages 680303–680303. SPIE, 2008. 3

[20] J. Kopf, M. F. Cohen, and R. Szeliski. First-person hyperlapse videos. In *SIGGRAPH*, pages 78:1–78:10. ACM, 2014. 3

[21] M. Lang, A. Hornung, O. Wang, S. Poulakos, A. Smolic, and M. Gross. Nonlinear disparity mapping for stereoscopic 3D. In *SIGGRAPH*, pages 75:1–75:10. ACM, 2010. 2, 3, 6, 8

[22] H.-S. Lin, S.-H. Guan, C.-T. Lee, and M. Ouhyoung. Stereoscopic 3D experience optimization using cropping and warping. In *SIGGRAPH Asia Sketches*, pages 40:1–40:2. ACM, 2011. 3

[23] W. Matusik and H. Pfister. 3D TV: A scalable system for real-time acquisition, transmission, and autostereoscopic display of dynamic scenes. In *SIGGRAPH*, pages 814–824. ACM, 2004. 3

[24] L. McMillan and G. Bishop. Plenoptic modeling: An image-based rendering system. In *SIGGRAPH*, pages 39–46. ACM, 1995. 3

[25] B. Mendiburu. *3D Movie Making: Stereoscopic Digital Cinema from Script to Screen*. Focal press, 2009. 1, 2, 3

[26] S. Moezzi, L.-C. Tai, and P. Gerard. Virtual view generation for 3D digital video. *Multimedia*, 4(1):18–26, 1997. 3

[27] B. S. Morse, J. Howard, S. Cohen, and B. L. Price. Patchmatch-based content completion of stereo image pairs. In *3DIMPVT*, 2012. 8

[28] R. Neumann. The lion king 3D: in-depth with disney. http://www.fxguide.com/featured/the-lion-king-3d-in-depth-with-disney/, 2011. [Online; accessed 14-August-2015]. 5

[29] K.-J. Oh, S. Yea, and Y.-S. Ho. Hole filling method using depth based in-painting for view synthesis in free viewpoint television and 3D video. In *Picture Coding Symposium, 2009. PCS 2009*, pages 1–4. IEEE, 2009. 8

[30] D. Pinskiy, J. Longson, P. Kristof, E. Goldberg, and R. Neuman. Stereo compositing accelerated by quadtree structures in piecewise linear and curvilinear spaces. In *Symposium on Digital Production*, pages 13–20. ACM, 2013. 3

[31] F. Pitié, G. Baugh, and J. Helms. Depthartist: a stereoscopic 3D conversion tool for CG animation. In *European Conference on Visual Media Production*, pages 32–39. ACM, 2012. 2, 3, 6

[32] S. Pujades and F. Devernay. Viewpoint interpolation: Direct and variational methods. In *International Conference on Image Processing*, pages 5407–5411. IEEE, 2014. 7

[33] S. Pujades, F. Devernay, and B. Goldluecke. Bayesian view synthesis and image-based rendering principles. In *Conference on Computer Vision and Pattern Recognition*, pages 3906–3913. IEEE, 2014. 3, 4, 6, 7, 8

[34] D. Scharstein and R. Szeliski. A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. *International Journal of Computer Vision*, 47(1-3):7–42, 2002. 3

[35] P. Seuntiens, L. Meesters, and W. Ijsselsteijn. Perceived quality of compressed stereoscopic images: Effects of symmetric and asymmetric jpeg coding and camera separation. *Transactions on Applied Perception*, 3(2):95–109, 2006. 8

[36] A. Shamir and O. Sorkine. Visual media retargeting. In *SIGGRAPH ASIA Courses*, pages 11:1–11:13. ACM, 2009. 3

[37] T. Shibata, J. Kim, D. M. Hoffman, and M. S. Banks. The zone of comfort: Predicting visual discomfort with stereo displays. *Journal of vision*, 11(8):11, 2011. 1, 2

[38] A. Smolic, H. Kimata, and A. Vetro. Development of MPEG standards for 3D and free viewpoint video. In *Three-Dimensional TV, Video, and Display IV*, pages 60160R–60160R–12. SPIE, 2005. 3

[39] A. Smolic and D. McCutchen. Report on 3DAV exploration of video-based rendering technology in MPEG. *Transactions on Circuits and Systems for Video Technology*, 14(3):348–356, 2004. 3

[40] R. Spottiswoode, N. L. Spottiswoode, and C. Smith. Basic principles of the three-dimensional film. *Journal of the Society of Motion Picture and Television Engineers*, 59(4):249–286, 1952. 1, 2

[41] M. Tanimoto. FTV (free-viewpoint television. *Transactions on Signal and Information Processing*, 1(e4):454–461, 2012. 3

[42] I. Tsubaki and K. Iwauchi. Depth remapping using seam carving for depth image based rendering. In *Image Processing: Algorithms and Systems XIII*, pages 93990R–93990R–12. SPIE, 2015. 3

[43] P. Vangorp, G. Chaurasia, P.-Y. Laffont, R. W. Fleming, and G. Drettakis. Perception of visual artifacts in image-based rendering of façades. *Computer Graphics Forum*, 30(4):1241–1250, 2011. 3

[44] Y.-S. Wang, C.-L. Tai, O. Sorkine, and T.-Y. Lee. Optimized scale-and-stretch for image resizing. In *SIGGRAPH Asia*, pages 118:1–118:8. ACM, 2008. 3

[45] M. Woo, J. Neider, T. Davis, and D. Shreiner. *OpenGL Programming Guide: The Official Guide to Learning OpenGL, Version 1.2*. Addison-Wesley Longman Publishing Co., Inc., 3rd edition, 1999. 6

[46] H. Yamanoue, M. Okui, and F. Okano. Geometrical analysis of puppet-theater and cardboard effects in stereoscopic HDTV images. *Transactions on Circuits and Systems for Video Technology*, 16(6):744–752, 2006. 1

[47] T. Yan, R. W. Lau, Y. Xu, and L. Huang. Depth mapping for stereoscopic videos. *International Journal of Computer Vision*, 102(1-3):293–307, 2013. 3, 8

[48] S. Zinger, L. Do, and P. de With. Free-viewpoint depth image based rendering. *Journal of Visual Communication and Image Representation*, 21(5):533–541, 2010. 3